EXHIBIT 4

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

AUTHORS GUILD, et al., Plaintiffs, v. OPEN AI INC., et al., Defendants.	**ECF CASE** No. 1:23-cv-08292-SHS; No. 1:23-cv-10211-SHS **PLAINTIFFS' THIRD SET OF INTERROGATORIES TO OPENAI**
JONATHAN ALTER, et al., Plaintiffs, v. OPENAI, INC., et al., Defendants.	

PLAINTIFFS' THIRD SET OF INTERROGATORIES TO OPENAI

Plaintiffs, by and through their undersigned attorneys, request that Defendants provide an answer to the Interrogatories within thirty (30) days of the date of service hereof as provided by Federal Rule of Civil Procedure 33.

DEFINITIONS

1. **"Communication(s)"** means the transmittal of information (in the form of facts, ideas, inquiries or otherwise) by any means, including, but not limited to, telephone calls, emails (whether via company server or personal webmail or similar accounts), faxes, text messages (on work or personal phones), instant messages, Skype, Line, WhatsApp, WeChat, other electronic messages, letters, notes, and voicemails.

2. **"Document(s)"** is defined to be synonymous in meaning and equal in scope to the usage of the term "documents or electronically stored information" in Rule 34(a)(1)(A).

3. **"You", "Your",** and **"OpenAI"** means OpenAI, Inc., OpenAI GP, LLC, OpenAI, LLC, OpenAI OPCO LLC, OpenAI Global LLC, OAI Corporation, LLC, OpenAI Holdings, LLC, and any of their directors, officers, employees, partners, members, representatives, agents (including attorneys, accountants, consultants, investment advisors or bankers), and any other person acting or purporting to act on their behalf, as well as corporate parents, subsidiaries, affiliates, predecessor entities, successor entities, divisions, departments, groups, acquired entities, related entities, or any other entity acting or purporting to act on their behalf.

4. **"Large Language Model," "LLM," "AI Model(s)," "Generative AI system(s)," "model(s)," and "API Product(s)"** have the same meaning as they are used in YOUR letter to the Register of Copyrights and Director of the U.S. Copyright Office dated October 30, 2023, "Re: Notice of Inquiry and Request for Comment [Docket No. 2023-06]" and include all models listed or described in https://platform.openai.com/docs/models.

5. **"Person(s)"** means any individual or entity.

6. **"Fine Tune(d)," "Fine Tuning," "Pre-Train(ed)," "Pre-Training," "Train[ed]," and "Training,"** have the same meaning as the term is discussed in YOUR website materials and statements. *See e.g.*, https://platform.openai.com/docs/guides/fine-tuning ("OpenAI's text generation models have been pre-trained on a vast amount of text. To use the models effectively, we include instructions and sometimes several examples in a prompt. Using demonstrations to show how to perform a task is often called "few-shot learning." Fine-tuning improves on few-shot learning by training on many more examples than can fit in the prompt, letting you achieve better results on a wide number of tasks. Once a model has been fine-tuned,

you won't need to provide as many examples in the prompt.”); https://openai.com/index/language-unsupervised/ (“These results provide a convincing example that pairing supervised learning methods with unsupervised pre-training works very well; this is an idea that many have explored in the past, and we hope our result motivates further research into applying this idea on larger and more diverse datasets.”); and https://openai.com/index/how-should-ai-systems-behave/ (“The two main steps involved in building ChatGPT work as follows . . . First, we “pre-train” models by having them predict what comes next in a big dataset . . . Then, we “fine-tune” these models on a more narrow dataset that we carefully generate with human reviewers who follow guidelines that we provide them.”); https://openai.com/index/gpt-4-research/ (“Interestingly, the base pre-trained model is highly calibrated . . . Note that the model’s capabilities seem to come primarily from the pre-training process . . . ”); https://openai.com/index/text-and-code-embeddings-by-contrastive-pre-training/ (“In this work, we show that contrastive pre-training on unsupervised data at scale leads to high quality vector representations of text and code.”); https://openai.com/index/vpt/ (“We trained a neural network to play Minecraft by Video PreTraining (VPT) on a massive unlabeled video dataset of human Minecraft play, while using only a small amount of labeled contractor data. With fine-tuning, our model can learn to craft diamond tools . . .”).

INTERROGATORIES

INTERROGATORY NO. 10:

Identify every version of every LLM, Generative AI system, AI Model, and API Product (e.g., GPT-3, text-davinci-003, davinci-instruct-beta, GPT-3.5-Turbo, code-davinci-002, GPT-4, gpt-4-base, babbage-002, GPT-4o, text-embedding-ada-002, ada:ft-openai:copyrightv1-books1-40epochs-2022-10-04-08-53-35 etc.), both as it existed (1) after pre-training and before fine-tuning, (2) after fine-tuning, and (3) any fine-tuned versions, including all models referenced on

https://platform.openai.com/docs/models, in **OpenAI's** possession, custody, or control, or which was previously in **OpenAI's** possession, custody, or control, as well as the date on which each model began training, completed pre-training, began any fine-tuning, completed any fine-tuning, and was last modified, including, any internal designation or description used for internal tracking or reference, and excluding any image or video generation model that did not train on text.

INTERROGATORY NO. 11:

For each model identified in response to Interrogatory No. 10 (including, as necessary, any internal designation or description used for internal tracking or reference), identify those which were used in any version of ChatGPT or any other **OpenAI** product or service as well as the product or service the model was used for, including, for the avoidance of doubt, any discontinued products or services, and products or services which were or are available via any internal or external API, but excluding products or services directed towards image or video generation which did not use any text-based large pretrained model, and the dates on which the model was first and last used for the product or service in question.

INTERROGATORY NO. 12:

For each model identified in response to Interrogatory No. 10 and product or service identified in response to Interrogatory No. 11, identify the parameters, properties, or filters which may be manipulated (by **OpenAI** or by a user) that may impact each model's output, including but not limited to, e.g., the "temperature" setting of the model, the sampling methods / sampling techniques (e.g., top-p or top-k), frequency and presence penalties, max length / min length, stop sequences, repetition penalties, random seed(s), any content or subject matter filters, and any other such parameters, properties, or filters.

INTERROGATORY NO. 13:

Identify by title and author all audiobooks that **You** transcribed to text.

INTERROGATORY NO. 14:

Identify all models identified in response to Interrogatory No. 10 which trained on any version of the Common Crawl dataset, and for all models identified, identify the version of the Common Crawl dataset (in whole or in part) it was trained on.

INTERROGATORY NO. 15:

Identify the location on **OpenAI's** computer systems and name (including any internal designation or description used for internal tracking or reference) of any dataset—used to train any model identified in response to Interrogatory No. 10—that may contain books, and, for any such dataset that is no longer in **OpenAI's** possession, custody, or control due to deletion, the date of such deletion. For each identified dataset, identify all models identified in response to Interrogatory No. 10 that trained on that dataset, or otherwise incorporated in any way the weights produced from training on that dataset.

Dated: May 17, 2024

/s/Justin A. Nelson
Justin A. Nelson (*pro hac vice*)
Alejandra C. Salinas (*pro hac vice*)
SUSMAN GODFREY L.L.P.
1000 Louisiana Street, Suite 5100
Houston, TX 77002
Tel.: 713-651-9366
jnelson@susmangodfrey.com
asalinas@susmangodfrey.com

Rohit D. Nath (*pro hac vice*)
SUSMAN GODFREY L.L.P.
1900 Avenue of the Stars, Suite 1400
Los Angeles, California 90067
Tel.: 310-789-3100
rnath@susmangodfrey.com

J. Craig Smyser
SUSMAN GODFREY L.L.P.
1901 Avenue of the Americas, 32nd Floor

New York, New York 10019
Tel.: 212-336-8330
csmyser@susmangodfrey.com

/s/Rachel Geman
Rachel Geman
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
250 Hudson Street, 8th Floor
New York, New York 10013-1413
Tel.: 212-355-9500
rgeman@lchb.com

Reilly T. Stoler (*pro hac vice forthcoming*)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
275 Battery Street, 29th Floor
San Francisco, CA 94111-3339
Tel.: 415-956-1000
rstoler@lchb.com
ibenserg@lchb.com

Wesley Dozier (*pro hac vice*)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
222 2nd Avenue, Suite 1640
Nashville, TN 37201
Tel.: 615-313-9000
wdozier@lchb.com

/s/Scott J. Shoulder
Scott J. Shoulder
CeCe M. Cole
COWAN DEBAETS ABRAHAMS & SHEPPARD LLP
41 Madison Avenue, 38th Floor
New York, New York 10010
Tel.: 212-974-7474
ssholder@cdas.com
ccole@cdas@com

Attorney for Plaintiffs and the Proposed Class

CERTIFICATE OF SERVICE

I hereby certify that on May 17, 2024, a copy of the foregoing was served via electronic mail to all counsel of record in this matter.

/s/J. Craig Smyser
(Signature)